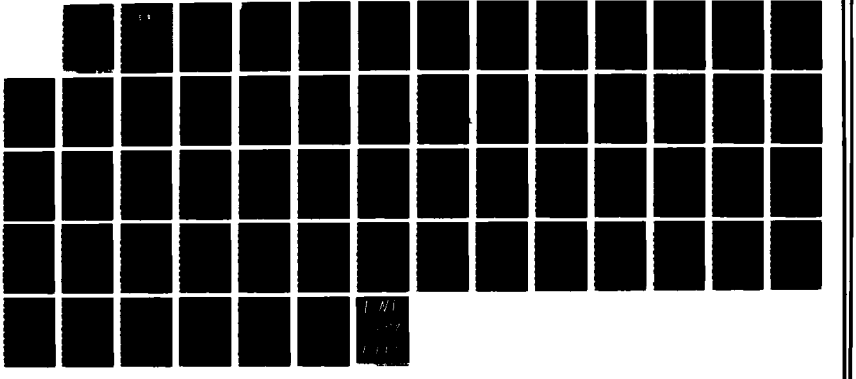
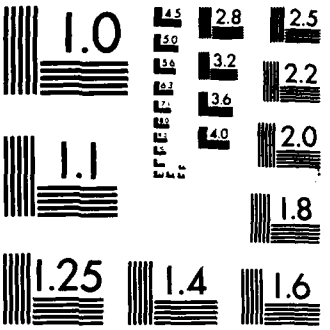
AD-A184 495 KINDS OF BOOTSTRAPS AND KINDS OF JACKKNIVES DISCUSSED IN TERMS OF A YEAR O (U) PRINCETON UNIV NJ DEPT OF STATISTICS J W TUKEY APR 87 ARO-23360 6-MA 1/1

UNCLASSIFIED DAAL03-86-K-0073 F/G 12/3 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARO 23360.6-MA
6a. NAME OF PERFORMING ORGANIZATION: Princeton University; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: U. S. Army Research Office
ADDRESS (City, State, and ZIP Code): Princeton, NJ 08544	7b. ADDRESS (City, State, and ZIP Code): P. O. Box 12211, Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: U. S. Army Research Office; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DAAL03-86-K-0073
8c. ADDRESS (City, State, and ZIP Code): P. O. Box 12211, Research Triangle Park, NC 27709-2211	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Kinds of Bootstraps and Kinds of Jackknives, Discussed in Terms of a Year of Weather-Related Data

12. PERSONAL AUTHOR(S)
John W. Tukey

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM _____ TO _____	April 1987	56

16. SUPPLEMENTARY NOTATION
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Weather-Related Data; Arithmetic Means; Resampling Methods; Bootstrapping; Resampling; Jackknifing

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

"Resampling methods" - - are in the process of becoming popular ways of assessing the standard error appropriate to some number chosen to distill, perhaps in a rather complex way, from data.

An ever-present danger is that "resampling" will come to be thought of as a cure-all. The most that can reasonably be hoped for is that questions that do

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

20. ABSTRACT CONTINUED

not arise in connection with the simplest distillates - - arithmetic means of samples - - need *not* be considered in connection with resampling applied to the results of more complicated calculations. (If, as we should be, we are content with reasonable approximations, this will often be the situation.)

This means that, in particular, (a) needs for the use of robust/resistant techniques in the distillation process, and (b) needs to consider the question of "the proper error term" are in *no* way automatically addressed by the use of jackknife or bootstrap. Robust/resistant techniques, if required, must be built into the calculation of the final distillate from the observations, before that distillate is jackknifed or bootstrapped.

A hypothetical, unspecified distillate from a year of weather-related data offers a good platform for careful discussion, one where the issues arise unavoidably and clearly, one where many of us because of all the years of weather we have lived through, have some "feel" for the facts. So we shall use this as an illustrative example - - one that raises many issues. We hope the modifications needed for other instances, whose essential character may be quite different, will be easy to make.

Our approach will proceed in four steps:

(A) the issues for arithmetic means,

(B) the issues for bootstrapping,

(C) the issues for jackknifing, and

(D) discussion, including comparison of jackknifing and bootstrapping.

Kinds of bootstraps and kinds of jackknives, discussed in terms of a year of weather-related data

by

John W. Tukey

Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544

Technical Report No. 292
Department of Statistics
Princeton University 08544

April 1987

Prepared in part in connection with research at Princeton University sponsored by the Army Research Office (Durham), DAAL03-86-K-0073.

TABLE OF CONTENTS

Kinds of bootstraps and kinds of jackknifes, discussed in terms of a year of weather-related data

John W. Tukey

Technical Report No. 292
Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544

1. Introduction.

"Resampling methods" - - are in the process of becoming popular ways of assessing the standard error appropriate to some number we have chosen to distill, perhaps in a rather complex way, from data.

An ever-present danger is that "resampling" will come to be thought of as a cure-all. The most that can reasonably be hoped for is that questions that do *not* arise in connection with the simplest distillates - - arithmetic means of samples - - need *not* be considered in connection with resampling applied to the results of more complicated calculations. (If, as we should be, we are content with reasonable approximations, this will often be the situation.)

This means that, in particular, (a) needs for the use of robust/resistant techniques in the distillation process, and (b) needs to consider the question of "the proper error term" are in *no* way automatically addressed by the use of jackknife or bootstrap. Robust/resistant techniques, if required, must be built into the calculation of the final distillate from the observations, before that distillate is jackknifed or bootstrapped. (While inadequately advertised, perhaps, this point has been made repeatedly.)

Prepared in part in connection with research at Princeton University sponsored by the Army Research Office (Durham), DAAL03-86-K-0073.

April 21, 1987

Questions related to "the proper error term" do not seem to have received appropriate attention. Their consideration will often determine the kind of bootstrap - - or, probably more often, the kind of jackknife - - required. This set of issues is the target of the present account.

We have referred to what we have chosen to calculate from the data as the "distillate". Some would have been happier had we used "estimate". The latter term usually implies, however:

a) a probability model (of greater or lesser tightness),

b) a parameter appearing in that model,

c) a desire to estimate that parameter,

d) a search for a "good estimate".

We do not want to rule this

probability model → parameter → estimate

route out. But neither do we want to rule out the

distillate → parameter → class of probability models

route, where the class of probability models is often vague. So we shall persist in using "distillate" for what is calculated from the data.

To balance things a little, we will use "parameter" for what our distillate is thought of as pointing to. (In the case of the second route, where the choice of the distillate has led to the choice of the parameter, some would insist that this usage is "par abus de langage" in Bourbaki's sense.) (We shall need this term only in Sections 18, 19 and 28.)

A hypothetical, unspecified distillate from a year of weather-related data offers a good platform for careful discussion; one where the issues arise unavoidably and clearly, one where many of us because of all the years of weather we have lived through, have some "feel" for the facts. So we shall use this as an illustrative example - - one that raises many issues. We hope the modifications needed for other instances, whose essential character may be quite different, will be easy to make.

Our approach will proceed in four steps:

(A) the issues for arithmetic means,

(B) the issues for bootstrapping,

(C) the issues for jackknifing,

(D) discussion, including comparison of jackknifing and bootstrapping

The writer sincerely thanks Bradley Efron for helpful comments and face to face discussions. He hopes to have represented Efron's views correctly, but must take the sole responsibility for all that appears here. He also thanks Thu Hoang and Kenneth W. Steinberg for help in reducing errors and improving clarity.

PART A: The issues for arithmetic means

The considerations of this part apply to many kinds of arithmetic means - - the mean temperature at 8 am for all the days of the year; the same for all hours of all days; the mean pollution of a specified sort, measured in a specified way, at a specified place, at 1 pm on Tuesdays throughout the year; the mean

temperature one hour after sunrise for all the days of the year - - and so on. For most of our considerations it will be sufficient to think of a very simple arithmetic mean, but our considerations apply quite generally to various kinds of arithmetic means.

Judging the variability of a distillate seems inevitably to be based on the identification of pairs of "parts" whose differences reflect, fairly and equitably, the impacts of those kinds of variation whose impacts on the chosen distillate we plan to account for. The simplest way to do this, not always feasible, is the single-block way, where the choice of error term is carried out in terms of *pieces,* where it is all the differences from one piece to another piece that are taken as the basis for assessing the chosen distillate's variability.

It may not be possible, especially with weather-related data, to use only a single block, as we shall discuss below. With several blocks each divided into pieces, the selected differences are all those between pairs of pieces that both belong to the same block.

We should think about blocks and pieces as exactly a way to identify differences that reflect the right sources of variability equitably.

2. Pieces and blocks.

Throughout our analysis we shall think of the observations as divided into pieces, which combine to make blocks. (A minimum of two pieces per block.) The underlying, almost qualitative, stochastic model is that we dare treat the pieces that make up any block as sample from a corresponding population of pieces and that we are going to treat the set of blocks as fixed. (Some would

then like to call "blocks" *strata*. (Those with a strong agricultural background might like to refer to "pieces" as "plots".)

When we come to resampling, we will operate block by block, in such a way as to preserve the sizes of the blocks.

In dealing with a year of weather-related data, our pieces are *not* likely to be *shorter* than a day, and are likely to be made up of (one or more) *whole* days. We might consider, for instance, 1-day, 3-day, 5-day or 10-day pieces and 10-day, 20-day or 1-month blocks.

An extremely naive approach to a year of weather-related observations would have chosen

pieces = 24-hour periods (few would dare go to 1-hour periods, even if the data would support it),

blocks = years (of which we have one).

This would have meant treating our 365 or 366 days as a random sample from a single population of days. This would be, as we shall see, a dangerously extreme choice in more than one direction, and is *not* what we shall recommend.

3. Why more than one day per piece?

Within a block we have to agree to dare to treat pieces as a sample. If 1-day pieces were to show appropriate behavior, so would k-day pieces, for larger k. The converse need not be true. What are the pros and cons of various choices of k? (We discuss these issues first for a single block, the extension to more blocks is easy - - and left to the reader - - also cp. Section 7.)

If we are dealing with a block of $n = km$ days of observations (see the top panel of exhibit 1), we can write our arithmetic mean over all the days of our year as

mean over m pieces

of the

mean over k days within a piece.

If we write, transiently, y_{jq} for the input from the j^{th} day of piece i, and put

$$\bar{y}_j = \frac{1}{k} \sum_q y_{jq}$$

$$\bar{\bar{y}} = \frac{1}{m} \sum_j \bar{y}_j$$

then the last equation leads to a Student's t with $m-1$ degrees of freedom and

$$s^2 = \frac{1}{m-1} \sum (\bar{y}_j - \bar{\bar{y}})^2$$

$$\text{est'd var } (\bar{\bar{y}}) = s^2/m$$

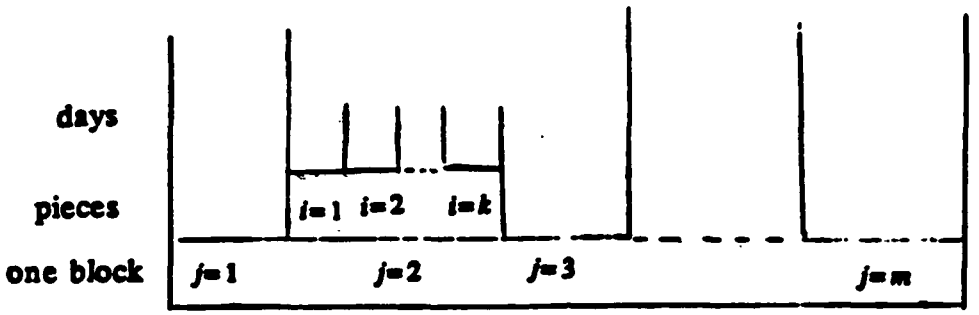
whatever the chosen value of k. In the ideal case we can choose k freely - - paying only a reduction to $m-1 = (n/k)-1$ degrees of freedom as k increases.

Real weather-related data tends to be persistent from one day to the next. Appreciable persistence, even of a probabilistic sort, extends for only a few days, perhaps 3 or 4. But adjacent day persistence is quite strong. The "meteorological events" of a weather sequence are not confined to individual days.

exhibit 1

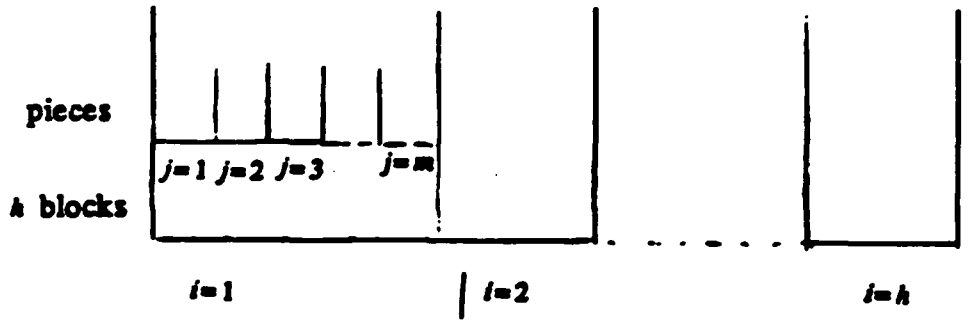
Days, pieces, blocks; relations and notation

As in Section 3:



(m pieces per block, k days per piece)

As in Section 7:



(m pieces per block)

April 21, 1987

The easy way to understand the direction of the resulting effect on assessing the variance of an arithmetic mean is to consider a very artificial example: So let us assume that n, the total number of days in our lone block, is even, and that days come in successive pairs such that the response from which we are distilling our arithmetic mean is the same for the two days of every twin pair, namely that

$$y_{2j-1} = y_{2j} \qquad \text{for } i \text{ from 1 to } n/2 ,$$

While the value for each twin pair is independent of all others. (This gives 50% persistence for adjacent days, and 0% persistence for days further removed from one another.)

If we use 2-day pieces that are in phase with the twin pairs, we have

$$\bar{y}_j = \frac{1}{2}(y_{2j-1} + y_{2j}) \qquad \text{for } j \text{ from 1 to } n/2$$

Since the $\bar{y}_j$ $(=y_{2j})$ *are* independent, a Student's t based on the variability among the $\bar{y}_j$, with its $(n/2)-1$ degrees of freedom, is an appropriate assessment (given all our assumptions) of the variability of $\bar{\bar{y}}$.

For this case, where $k=2$, each pair of days, presently one piece, contributes $(y_{2j-1}-\bar{\bar{y}})^2 = (y_{2j}-\bar{\bar{y}})^2$ to the sum of squared deviations - - since $\bar{y}$ (for the j^{th} pair) $= y_{2j-1} = y_{2j}$ - - this sum would then be divided by $(n/2)-1$ to reach an $s^2_{[2]}$, where [2] reminds us of the piece size. If, instead, we had taken $k=1$, each pair of days, now two pieces, would contribute $(y_{2j-1}-\bar{\bar{y}})^2 + (y_{2j}-\bar{\bar{y}})^2 = 2(y_{2j}-\bar{\bar{y}})$ to the new sum of squares, which would consequently be twice what it was for $k=2$. This sum is then divided by $n-1 = 2((n/2)-(1/2))$ which is only a little larger than $2((n/2)-1)$. Thus, in this very special

circumstance, $s^2_{[1]}$, the s^2 for pieces made up of single days, is only slightly smaller than $s^2_{[2]}$, the s^2 for pieces made up of two twin days.

If we try to get away with using $k=1$ to reach an estimated variance for $\bar{\bar{y}}$, we will divide this slightly smaller $s^2_{[1]}$ by n, when we already know that, in our special case, dividing the slightly larger $s^2_{[2]}$ by $n/2$ gives an appropriate variance estimate for $\bar{\bar{y}}$. Thus, in this special case, using $k=1$ leads to an estimated variance only about 50% of what is appropriate.

If there is strong persistence, we dare not use $k=1$!

To carry this special case a little further, we return to $k=2$ and suppose that the 2-day pieces straddle the boundaries between the twin pairs. Adjacent $\bar{y}_i$ will now be correlated, with $r = .5$ and $r^2 = .25$, while more remote $\bar{y}_i$ will remain uncorrelated. The amount of distortion corresponds, roughly to:

k = 1: half adjacent pairs (same twins): $r^2 = 1$

other adjacent pairs (straddle): $r^2 = 0$

k = 2 pieces out of phase with twin pairs: $r^2 = .25$

pieces in phase with twin pairs: $r^2 = 0$

Since, in a real situation, such a regular pattern of dependence would *not* persist, it is reasonable to average the match-straddle alternatives as well as the two kinds of adjacency, giving, for this special case:

$k = 1$: $\bar{r^2} = .5$ $(|\bar{r}| = .5)$
$k = 2$: $\bar{r^2} = .125$ $(|\bar{r}| = .25)$

This indicates, for cases like this extreme case, that going to $k = 2$ should

greatly ameliorate the bias due to persistence, but that going to $k = 3$ or more, which will reduce r^2 further, is likely to be worthwhile.

Unless there is some unusually stringent reason based on other considerations, then, we ought *never* take $k = 1$ in dealing with arithmetic means of weather-related quantities, and we *ought to look forward* to $k = 3$ or more. (We expect this to generalize to other distillates.)

4. Why more than one block per year?

Every year of our present calendar has 31 January days and 31 July days. However, if we think of days as a sample from a single population of days (which surely can be thought of as retaining their month-name tags) a simple sample would have about

$$31 \pm 2\sqrt{31} \quad (\pm 2\sigma\text{-limits run from 20 to 42})$$

days with January tags and, similarly, 20 to 42 (at $\pm$ 2 σ) July-tagged days. This is most unlikely to be acceptable, especially anywhere outside the tropics, where July days are very different from January days.

If we make 12 blocks, one for each month, we will ensure exactly 31 July days and exactly 31 January days in each resample. But this may not be enough. In northern temperate climates, March traditionally "comes in like a lion and goes out like a lamb". Of our 31 March days $15 \pm \sqrt{7.5}$, (9.5 to 20.5 at $\pm$ 2 σ) will be lion-like 1 March to 15 March days. Are we prepared to accept this much variation in lion-like March days? Often not! So we need to be prepared to use blocks shorter than one month, at least for those times of year when the seasons change most rapidly.

5. Why not really small blocks?

Should we go to the other extreme, and use very short blocks, like 2-day or 4-day blocks? (These would require one-day or two-day pieces.) Almost surely not, for if we take our blocks short enough for a single meteorological event (we use this term for the time involved in a single weather system - - not just a single hour or a single day) to cover that block, every piece in this block will involve the one event, and we are forcing - - in one sense or another - - this event to be represented in *all* of our alternatives whether merely as contributors to σ^2 or in some kind of resamples. This is unrealistic! There will be 31 days in July next year, but a copy next year of this year's most unusual (short-term) event is quite unlikely.

So we dare not take our blocks too short.

6. Longer-term weather phenomena.

The weather patterns we all are used to noticing last for a few days at a time. Underneath these alternations there are longer-term irregular changes of smaller magnitude. Sea-surface temperatures in the central Pacific seem to have season-long effects on North American weather. So-called "blocking events" can last for weeks or even months. There are differences from one year to another that cannot be assessed from what goes on within a single year.

Often we would like to include the consequences of these year-to-year variations in our "significance" or "confidence" statements about our results - - sometimes it would be quite essential to do this - - but if we have only one year's data we just cannot do anything in the way of relevant calculation. All

we can do is to emphasize that, for example, our confidence interval is based on only a year's data and that, accordingly, the interval needs to be widened by a judgment-based amount to allow for unmeasured, year-to-year variation.

7. Student's t for blocks.

If we have h blocks each divided into m pieces, we are dealing with hm results $\{y_{ij}\}$, for individual pieces, where i runs over blocks from 1 to h and j runs over pieces from 1 to m. For the i^{th} block

$$\bar{y}_i = \frac{1}{m}\sum_j y_{ij}$$

$$s_i^2 = \frac{1}{m-1}\sum_j (y_{ij} - \bar{y}_i)^2$$

$$\text{est'd var}\{\bar{y}_i\} = \frac{s_i^2}{m}$$

combining over blocks gives, since we dealing with independently estimated $\bar{y}_i$.

1) $\bar{\bar{y}} = \frac{1}{h}\sum_i \bar{y}_i$

(*) $\text{est'd var}\{\bar{\bar{y}}\} = \frac{1}{h^2}\ \text{est'd var}\{\sum \bar{y}_i\} = \frac{1}{h^2}\sum\ \text{est'd var}\{\bar{y}_i\} \frac{1}{h^2 m}\sum_i s_i^2 = s^2/hm$

2) where $s^2 = \frac{1}{h}\sum_i s_i^2$.

Thus the natural form for Student's t is

3) $t = \dfrac{\bar{\bar{y}} - \text{its contemplated value}}{\sqrt{s^2/hm}}$

as expanded by (1) and (2).

8. Some specific possibilities.

For our year of data, there are a number of more or less plausible choices

of piece size and block size, including:

— pieces of	—— days each	per block of —— days	(df)
2	4	8	(45)
3	3	9	(80)
2	5	10	(36)
3	4	12	(60)
4	3	12	(90)
5	3	15	(96)
3	5	15	(48)
6	3	18	(100)
4	5	20	(54)
5	4	20	(72)
2	6	12	(30)
2	8	16	(23)

each of which has $k \geq 3$. (The last column gives degrees of freedom accumulated, optimistically - - see Section 10 below - - for a year.) We can also consider some (less desirable) cases with $k = 2$, mainly:

—— pieces of 2 days each	per block of —— days	(df)
4	8	(135)
5	10	(144)
6	12	(150)
8	16	(154)
10	20	(162)

which could provide some extra degrees of freedom, even though they may not be effective enough in dealing with persistence to estimate a large enough s^2 to be appropriate.

9. Composites of arithmetic means.

Often our attention needs to be given to results that compound (perhaps by taking the difference) different arithmetic means. In the simplest case, as when the difference in arithmetic means for Chicago and Milwaukee is at issue,

we can arrange to t-test the final distillate in a way that bypasses any attempt (a) to consider an assumed amount of correlation between intermediate results, or (b) to evaluate the variability of the sampling distributions of the separate arithmetic means before differencing. (Doing (b) would of course force us to do (a)). (Assuming no correlation is often even more dangerous than assuming a specific amount, e.g. that $r = .42$.) When we can work directly with the final distillate (as we can both in this example and in almost all resampling situations), it will almost always be wiser to do so.

10. Seasonality of variability - - and degrees of freedom.

We need to think somewhat further about our final s^2, in particular about assigning it an appropriate number of degrees of freedom. To think effectively about this topic is easier if we recall "df" might better have been called "degrees of firmness" and that what we protect against by using, for instance, the numerically larger 5%-points of Student's t for finite values of ν (larger as compared with the Gaussian value of 1.960 for $\nu = \infty$) is the chance that s^2, in a particular instance, will be substantially smaller than the σ^2 it is estimating.

If the variability of whatever we are averaging is different from one season to another, as would be the case with average daily snowfall in middle latitudes, for instance, we need to allow for this fact in assessing a df for our s^2 . In the snowfall case, for example, there will be no variability in summer. Adding the s_i^2 for the summer blocks does nothing to improve the coefficient of variation of s^2 = (sum of the s_i^2)/h .

If we want to assign a reasonable number of degrees of freedom to s^2 or s^2/h (which deserve the same number), the simplest thing that we know how to do is the Smith-Welch-Fisher calculation of an equivalent number of degrees of freedom, ν_{eq}:

$$\frac{1}{\nu_{eq}} = \sum_i \left[\frac{s_i^2}{s^2 + \ldots + s_h^2} \right]^2 \frac{1}{\nu_i}$$

where s_i^2, for the i^{th} block, deserved ν_i degrees of freedom.

If the individual s_i^2 have only a few degrees of freedom, particularly when they have only one or two, the SWF analysis may not be satisfactory.

The reason is simple. If the σ_i^2 were all equal, as were the ν_i, then

$$\sum \left[\frac{\sigma_i^2}{\sigma_1^2 + \ldots + \sigma_n^2} \right]^2 \frac{1}{\nu_i} = \sum \left(\frac{1}{n}\right)^2 \frac{1}{\nu} = \frac{1}{n} \frac{1}{\nu}$$

giving $\nu_{eq} = n\nu$. But if the ν_i are small the values of the s_i^2 will not be close to $\sigma_i^2 = \sigma^2$ and the SWF formula, using the s_i^2 we have instead of the σ_i^2 we don't, will give too few degrees of freedom. The values of an untypically typical (untypically close to ideal behavior) set of 10 s_i^2 with $\sigma_i^2 = 1$ and $\nu = 2$ might well run from 2.74 and 1.82 down to .18 and .07. Putting such a set of 10 values in the SWF formula could easily give $\nu_{eq} = 11.56$ rather than the $\nu_{eq} = 20$ that we deserve. (For $\nu_i = 4$ we might get $\nu_{eq} = 29$ instead of 40. For $\nu_i = 1$ we might get 4.23 instead of 10.)

Even a rough fix for this problem can be worthwhile. What we recommended (provided the ν_i are nearly equal) is to:

1) order the s_i^2 and renumber them so that $s_1^2 \le s_2^2 \le \ldots \le s_n^2$ (relabel the ν_i accordingly)

2) let $a(i \mid n, \nu)$ be an approximate median for the i^{th} order statistic in a sample of n from χ^2_ν/ν,

3) use the Wilson-Hilferty approximation and the usual working-value approximation for Gaussian order statistics to write

$$a(i \mid n, \nu) \doteq (1 - \frac{2}{9\nu} + a(i \mid n)\sqrt{(2/9\nu)}\,)^3$$

where

$$a(i \mid n) = Gau^{-1}(\frac{3i-1}{3n+1})$$

4) divide each s_i^2 by $a(i \mid n, \nu_i)$,

5) apply SWF to the resulting ratios.

For $\nu = 2$, we know that $a(i \mid n, \nu)$ can be found from

$$1 - e^{-a(i \mid n, 2)} = \frac{3i-1}{3n+1}$$

and hence from

$$a(i \mid n, 2) = -\ln(1 - \frac{3i-1}{3n+1})$$

avoiding the use of the Wilson-Hilferty approximation for this particular value of ν.

While the Wilson-Hilferty approximation is usually recommended for large ν, where it is indeed very precise, it works surprisingly well for small ν. The writer would not mind, in the present context where even a rough correction is very worthwhile, using it for $\nu = 1$ or $\nu = 3$, as well as for intermediate values of ν.

We may also need to think, in judging how much data we really have, about effective - - or relevant - - numbers of blocks. Other forms of

calculation, some of which depend specifically on the distillate chosen, are likely to be appropriate.

11. Summary of PART A.

This part has been concerned with the basic questions about "choice of error term" for the simplest, relatively highly manageable distillates, namely arithmetic means. The two basic questions were:

1) what comparisons should contribute to our s^2 ?

2) how many degrees of freedom do we deserve?

The three basic answers turned out to be:

1) Comparisons of days not too near one another (not adjacent days) and not too far from one another (not July with January).

2) We can only earn contributions to degrees of freedom form a particular blocks if that block contributes appreciable variation to the final distillate,

3) we need to assess an equivalent number of degrees of freedom, plausibly by the Smith-Welch-Fisher *calculation*, often after adjustment for sampling dispersion.

These results apply to a wide variety of arithmetic means, as illustrated at the start of this part. (With the appropriate modifications to the formulas for Student's t, they would also apply to weighted arithmetic means.) We have no reason not to expect them to apply to many other distillates.

This being the situation for the simplest distillates, the best we can hope for from resampling is that, when we use resampling wisely enough, there will

be no new kinds of problem for at least moderately complex distillates.

PART B. Bootstrap Issues

12. The basic bootstrap.

As described by Bradley Efron, the basic bootstrap operates by treating an observed set of observations as a (finite) population of pieces, and sampling pieces from it with replacement in such a way as to generate resamples of the same size as the original sample.

This means beginning by converting the finite sample that was observed, not into a *finite* population as jackknives do, but rather into an infinite population where the only values present are those actually observed in the sample and the discrete probabilities of such values are the discrete frequencies observed in the data. If the data consists of only two pieces, A and B, - - not too likely to happen, but the easiest case to describe in detail - - then:

25% of resamples will be made up of two copies of A

50% will consist of one A and one B, like the original

25% will consist of two copies of B

Usually there will be many more than two pieces. "Resamples" or "bootstraps" are drawn, using the best available pseudo-random numbers. What would seem to be enough "resamples"?

To the extent that Student's t reflects our knowledge about the original distillate, if the actual variability deserved v degrees of freedom - - ordinarily

$\nu \leq n-1$, where there were n original pieces - - and we draw N bootstraps, the bootstrapped variability might deserve (roughly) only about $\nu^*(< \nu)$ degrees of freedom, where

$$\frac{1}{\nu^*} \approx \frac{1}{\nu} + \frac{1}{N}$$

(or perhaps $\approx (1/\nu) + (c/N)$ for some moderate c). If ν is small, $N \geq 10\nu$ provides most of the information possible. If ν is large, $N \geq \nu$ makes ν^* quite large, so that "larger" would not be appreciably better. Thus 100 to a few hundred bootstraps ordinarily suffice.

13. The blocked - - or stratified - - bootstrap.

Here, our resampling is restricted to selecting the correct number of pieces from each block (or stratum), again with replacement. If we have pieces A and B in block 1 and pieces C and D in block 2, then the sampling we strive for will have:

6.25%	AACC's
12.50%	ABCC's
6.25%	BBCC's
12.50%	AACD's
25.00%	ABCD's
12.50%	BBCD's
6.25%	AADD's
12.50%	ABDD's
6.25%	BBDD's
100%	

with similar situations - - showing independence from block to block - - for larger examples.

14. Duplication in the bootstrap.

Our examples, so far with (a usually quite unrealistic) 2 pieces per block, have made it clear that some duplication of pieces are to be expected. How many times will a particular piece appear in a given bootstrap when there are m pieces per block? Exhibit 1, easily calculated from the formulas for binomial distributions, has the answers.

Contrary to the intuition of many, the problem of duplication is least for 2 pieces per block, which has only 25% omissions, a fraction that rises steadily as k increases. The fraction of repetition twice or more, also begins at 25%, but rises only very slightly, remaining below 27%.

15. Theoretical results.

Derivations for the bootstrap are usually confined to limiting (asymptotic) results as the number of pieces per block get larger and larger. (Just doing many, many bootstraps - - 500, 5000, even 50,000 - - provides no added applicability for the classical derivations of the bootstrap.)

The theory of many-block bootstraps seems not to have been worked through.

As I have heard Efron say, the bootstrap was invented, using the jackknife as a model, to be supported by a simpler and more coherent theory and thereby support the jackknive. This has meant that it has had, to a degree not always publicized, to be a "large sample" (which here means "many pieces per block") technique. The basic asymptotic results involve terms of order $1/\sqrt{n}$, and are equally applicable to the bootstrap and to a wide variety of modifications of the

exhibit 1

Frequency of multiple appearances for any given piece,
when there are k pieces per block*

appearances	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	k large
0	25%	30%	33%	33%	37%	37%
1	50%	44%	42%	41%	40%	37%
2	25%	22%	21%	20%	20%	18%
3		4%	5%	5%	5%	6%
≥ 4			0.4%	0.7%	0.9%	1.9%
(2 or more)	(.250)	(.259)	(.262)	(.263)	(.263)	(.264)

*Values shown are, respectively, 1, 2 and 1 divided by 4;

8, 12, 6 and 1 divided by 27;

81,108, 54, 12 and 1 divided by 256;

1024, 1280, 640, 160, and 21 divided by 3125;

15625, 18750, 9375, 2500 and 406 divided by 416,656;

and 1, 1, 1/2, 1/6, $e-(8/3)$ divided by e

(the last from the Poisson approximation to the binomial).

bootstrap (including the classical versions of the jackknife). All that is needed is the effects of the modifications converge (to zero) even slightly more rapidly than $n^{-1/2}$ as $n \to \infty$.

It would be most helpful if we could make even rather crude estimates of how much impact on the bootstrap the finiteness of n is having in specific instances.

One simplicity of the bootstrap is the idea of using the 2.5% points of the distribution of bootstrap results as a 95% confidence interval for the final result - - a very attractive (or seductive) approach. Even for sufficiently large n, this

has problems (see Section 19, below.) But if the experience with the jackknife is a good analogy, then, as discussed later, we cannot expect this to work at all well for small n - - for a few pieces per block. The most that we can hope for is that we can use the s^2 of the bootstrap results to estimate the variance of the lone actual result, which means that we will have to depend, as is so often the case, on the robustness of Student's t.

If, as our discussion above indicates to be the case for weather-related data, we want to work with relatively few pieces per block, dare we bootstrap such blocks?

If we had no other alternative, the sound maxim that "It is better to have an approximate answer to the correct question, rather than an exact answer to the wrong one!" would tell us "Don't bootstrap with one block per year, that's quite unacceptable, it's better to bootstrap with small blocks even though we do not know as much about the precision (or even the accuracy) of our final results as we should!" However, as the remaining parts will make clear, we do have other alternatives.

16. The correction for finiteness.

As Efron has pointed out to me, the conversion of a sample of m into an infinite population (with m equal parts) changes the variance, since we now divide by m where we did divide by $m-1$. The variance of the results of bootstrap resampling is thus only $(m-1)/m$ times as large as would be needed to reflect the original sampling variance.

This effect is of order $1/m$, while the asymptotic calculations are only correct to order $1/\sqrt{m}$. Thus it is reasonable to neglect the finiteness correction for large m (where it will be small).

The writer, however, sees no excuse for neglecting this correction for small m. It corresponds to a phenomenon that we know happens for the simplest instances of distillates. It is reasonable to take it as a useful approximation for other distillates.

Thus, any calculation using the variance of some quantity over bootstraps resamples, the writer would be careful to multiply by $m/(m-1)$, which is as large as 2 for $m=2$, before inserting the result in the appropriate formula.

When we have several blocks, $1/m$ will be much larger than $1/mh$, quite possibly as large as $1/\sqrt{mh}$ which is about the order to which we might have a blocked bootstrap be asymptotically correct. Here there seems absolutely no excuse for avoiding the correction for finiteness.

17. The extreme cases of the ultimate and of the penultimate.

An extreme case that challenges the bootstrap very effectively arises when the distillate is the largest (or smallest) value of some observable quantity during the year observed. If $z_1 \geq z_2 \geq z_3 \geq \ldots$ are the largest values in the different pieces concerned, then, in unblocked bootstraps involving many pieces

z_1 will be the maximum in about 63% (= 100%-37%) of all cases

z_2 will be the maximum in about 23% (63% of 37%)

z_3 will be the maximum in about 9% (63% of 14%)

z_4 will be the maximum in about 3% (63% of 5%) and so on.

The first of these for instance, has to happen because, see Section 14, there is a 37% chance of any specific piece being omitted from any specific bootstrap shape. In fact, the chance of any one piece being missing is about 37%, with reasonable independence for different pieces.

This means that if we try the seductive approach of looking at the 2.5% points of the distribution of bootstrap results, enough repetitions will give z_4 and z_1 as, notionally, the 95% confidence limit for "the largest". Clearly such a confidence interval is quite unsatisfactory, since a confidence interval pretends to tell us about some long-run value, some parameter. In fact, the median (over years) largest (in a given year) will be larger than z_1 50% of the time. (The mean largest, for the usual behavior of distribution tails, will be still larger than the median. Thus, it will exceed z_1 *more* than 50% of the time.) Against such an extreme challenge, this seductive approach fails lamentably.

It might be that the s^2 based on bootstrap repetitions is not too bad an estimate of the variance of z_1, but it would be surprising if the adequacy of this approximation did not depend upon the shape of the distribution tail, just as the amount of asymmetry of the distribution of z_1 (and thus the severity of the corresponding challenge to Student's t if we try to use $z_1 \pm t_0 s$ as a confidence interval) surely does. Again, if we had no alternative - -

The second highest value is not as great a challenge as the highest one, but it is still a very serious challenge. It will simplify our analysis, and sometimes be reasonable, to assume that the 5 or 6 highest values fall in different pieces. (This will surely be the case when the values apply to pieces, and not to days or hours within them.)

Now we have the following event-descriptions for a bootstrap:

z_1 will be second highest if piece 1 occurs at least twice

z_2 will be second highest if piece 1 occurs once and piece 2 at least once OR if z_1 is absent and z_2 appears at least twice

z_3 will be second highest if z_1 and z_2 appears a total of only once and z_3 appears at least once, OR if z_1 and z_2 are both absent, and z_3 appears at least twice

z_4 will be second highest if z_1, z_2 and z_3 appear a total of once and z_4 appears at least once OR if z_1, z_2, z_3 are all absent, and z_4 appears at least twice.

The relevant individual-piece probabilities are, for many pieces,

absent: 36.8%

just one: 36.8%

at least once: 63.2%

at least twice: 26.4%

and the compound probabilities are:

two a total of once: $(36.8\%)(36.8\%) + (36.8\%)(36.8\%) = 27.0\%$

two not at all: $(36.8\%)(36.8\%) = 13.5\%$

three a total of once: $3(36.8\%)^3 = 14.9\%$

three not at all: $(36.8\%)^3 = 5.0\%$

April 21, 1987

from which we get the following probabilities that z_1 will be the 2nd highest in a bootstrap repetition.

i	Probability z_i is 2nd highest			
1				26.4%
2	(36.8%)(63.2%)	+	(36.8%)(26.4%)	= 33.0%
3	(27.0%)(63.2%)	+	(13.5%)(26.4%)	= 20.6%
4	(14.9%)(63.2%)	+	(5%)(26.4%)	= 10.7%
5	(7.3%)(63.2%)	+	(1.8%)(26.4%)	= 5.1%
6	(3.4%)(63.2%)	+	(0.7%)(26.4%)	= 2.3%
(≥ 7)				(1.9%)

These probabilities may not be unreasonable, but we have still to worry about the impact of different kinds of tail area shapes.

Were there no alternatives, we might have to do such things, but . . .

We notice how silly it would be to calculate actual bootstrap replications in each of these special, unusually challenging situations, since we can calculate the results for all replications rather easily.

18. The reflection of dissymmetry.

* naive argument *

Suppose that the true distribution of some quantity in samples of n is quite skew, say with a stretched tail to the right, and that n is large enough so that we are supposed to trust the bootstrap procedure. This means, presumably, that our bootstrap samples will have a distribution close to the actual sampling distribution - - and will also be stretch-tailed to the right.

What ought our confidence limit do? The danger that the observed value is far to the left of the value of the parameter is much smaller than the danger that the stretched right tail has let it go far to the right.

Thus our confidence interval should be longer to the left (thus allowing for the observed value to have been further to the right) than to the right. This is the opposite of the true sampling behavior's distribution. If the bootstrap distribution is like the sampling distribution, what we need is also the opposite of the bootstrap distribution's behavior.

It seems hard for the direct use of the bootstrap distribution to allow correctly for the shape of the sampling distribution. What the bootstrap is more likely to do well is to assess the variability of the sampling distribution.

* careful discussion *

The argument just given is correct for an important special case, when the possible states of the world differ by rigid translation. When we deal with one end of a confidence interval, the relevant theory is not what applies when the parameter falls at the value we observed, but rather that which applies when the parameter falls *at that end* of the confidence interval. Efron has pointed out that, in the most elementary of those cases where changing the parameter shrinks the distribution toward zero or swells up away from zero - - where we are dealing with a scale parameter - - the naive argument fails, because the confidence interval associated with knowing exactly what the possible distributions are is skewed in the *same direction* (but not by the same amount) as the sampling distribution. This might seem to considerably reduce the

cogency of the naive argument. However, as we next see, there is a better way to look at this class of examples.

If instead of using a distillate to tell us about a scale parameter τ, we choose to look for one to tell us about $\log \tau$, we will have converted a scale specification into a location specification, usually changing what was a heavy skew to the right into a somewhat milder skew to the left. (Doing our best to eliminate the skew often leads to focussing on something like τ^{-3}.)

If we are not going the seductive route, it *will* matter whether we focus on τ, $\log \tau$, or τ^{-3}. This discussion suggests that something between making the distribution nearly symmetric and making the effect of the parameter nearly additive can often lead to a better-posed problem.

* conclusion *

To go from the lone actual value, and a bootstrapped variance, to a confidence interval, then requires something like Student's t. If we do not require to include an exceptional degree of robust/resistance (within this final stage of setting a confidence interval - - which is a quite different issue from building in robustness/resistance into the calculation that distills the lone actual value out of the data), it is likely that Student's t is close to the best that we might do.

19. Less seductive and safer bootstraps.

Neither Efron or the writer recommend the use of the seductive bootstrap (in which the 2.5% points of the bootstrap distribution are taken as the ends of

a "95%" confidence interval). While we differ in the type of improvement we prefer, we are in agreement that improvement is needed.

For Efron's detailed suggestions, see Section 7 of his recent paper (Efron 1987). We notice here only that they include using % points for modified %'s, not just for equal tails.

The writer would, when bootstrappery is at issue, recommend at least trying:

1) seek for an expression for the parameter at interest for which the sampling distribution of the corresponding estimate (corresponding distillate) - - as pictured by the bootstrap distribution - - is reasonably symmetrical (about where it wants to be symmetrical, not necessarily about the value of the distillate based on the actual observations).

2) apply the seductive procedure to provide an initial confidence interval,

3) expand the length of this interval in the ratio $(m/(m-1))^{1/2}$, and recenter it at the distillate from the actual data.

If it doesn't seem feasible to do (1), I would consider

1') Seek for an expression for the parameter of interest that makes the changes in distribution consequent upon changing the parameter as much like translation as we can,

2') Like 2.

3') Like 3.

In any event, don't just sit there and use the seductive bootstrap.

April 21, 1987

20. Summary of PART B.

The basic idea of the bootstrap is quite simple. Its theoretical support is asymptotic - - for large samples, which for us means many pieces per block. Blocking, which seems essential for single years of weather-related data, tends to force us to small samples.

With fewer pieces per block, correction for finiteness is likely to be needed.

In any event, the seductive bootstrap (using the 2.5% points of the bootstrap distribution as the ends of a confidence interval) is to be avoided.

It is far easier for a resampling procedure to assess the *variability* of the final distillate than to assess the *shape of its distribution*. So even with the bootstrap we are likely to be driven to a symmetric interval for a re-expressed parameter, even, perhaps, to Student's t.

If we had no other alternative with a year of data, we would presumably decide to use a blocked bootstrap, theoretical uncertainties and all. However, as we shall soon see, we do have alternatives.

PART C: Jackknife Issues.

21. The basic jackknife.

If we have but one block, and m pieces, the basic (or leave-out-one) jackknife begins with $m+1$ distillations; one, y_{ALL}, using all the data, and m others $y_{(1)}, y_{(2)}, \ldots, y_{(m)}$ where $y_{(j)}$, read y-not-j is distilled from *all* the data *except* the j^{th} piece. The next step is to calculate m *pseudo-values* from

$$y_{\bullet j} = m y_{ALL} - (m-1) y_{(j)}$$

and the final step is to use Student's t "as if" the pseudo-values were a sample, through:

$$y_\bullet = \frac{1}{m}\sum y_{\bullet j}$$

$$s_\bullet^2 = \frac{1}{m-1}\sum (y_{\bullet j} - y_\bullet)^2$$

$$\text{est'd var } \{y_\bullet\} = s_\bullet^2/m$$

$$t = \frac{y_\bullet - \text{its contemplated value}}{\sqrt{s_\bullet^2/m}}$$

By using $y_\bullet$ instead of y_{ALL}, the jackknife makes a bias correction appropriate for biases proportional to 1/(amount of data). Since

$$y_\bullet - y_{ALL} = \frac{1}{m}\sum (y_{\bullet j} - y_{ALL}) = \frac{m-1}{m}\sum (y_{ALL} - y_{(j)})$$

$$= (m-1)(y_{ALL} - \bar{y}_{(\cdot)})$$

where

$$\bar{y}_{(\cdot)} = \frac{1}{m}\sum y_{(j)}$$

This bias correction, for a shift from $1/m$ to 0, is $m-1$ times the estimated change in bias from $1/(m-1)$ to $1/m$ (a distance of $1/m(m-1)$), as it should be.

Notice carefully the restriction to Student's t in connection with "as if the pseudo-values were a sample". Attempts to use other confidence procedures, like the sign test and the Wilcoxon test, for example, within the jackknife have proven unsatisfactory. It is easiest to remember the formulas in terms of "as if . . . a sample", but it is not wise to think about the procedure in such a way. Rather we should think of the jackknife as a way of producing a $y_\bullet$ and an $s_\bullet^2$ such that:

1) the average of $y_\bullet$ is what y_{ALL} seems to have been trying to estimate, and

2) we have near equality in

$$var\,(y_\bullet) \approx ave\,(s_\bullet^2/m)$$

so that a Student's t with $y_\bullet$ in the numerator and $\sqrt{s_\bullet^2/m}$ in the denominator is quite appropriate.

The more closely our distillates resemble arithmetic means in their behavior, the clearer it is that the jackknife process works well.

The most challenging situations for the jackknife (and presumably, also for the bootstrap) are those where the values associated with a very few of the pieces control the value of the distillate, where there is *narrow estimation*.

Narrow estimation is antipodally opposite to the behavior of arithmetic means. The most extreme (and most antipodal) form of narrow estimation arises when our distillate is - - or behaves rather like - - an *order statistic* (and, at least for simplicity, each piece contains exactly one value) - - for example, a minimum, quartile, median or maximum.

Among order statistics the behavior of an extreme of - - the maximum or the minimum - - is the furthest from the behavior of the arithmetic mean. This is one of the reasons why we began to discuss extremes and near-extremes in Section 17 of PART B.

Extremes and near-extremes are a severe challenge to the jackknife. Knowing this need not keep us from using the jackknife when we are sure that we should distill out an extreme. But knowing this should urge us to ask

ourselves: "Can we modify the distillation process to serve the same ends nearly as well so that the behavior of the distillate involved is not so extreme?".

If we have to deal with an intermediate order statistic, or with some other distillate that behaves rather similarly, the basic or "leave-out-one" jackknife has very limited effectiveness. The reason for this is easy to see. If we leave out any value below that of the distillate, we have one common effect on the order statistic. Similarly, for leaving out any value above the distillate we get a second common value for $y_{(i)}$. With at most 3 different values for the $y_{(i)}$'s (we could also leave out the distillate (the original order statistic) itself!), there can be at most 3 different values for the y_{*i}. So how can the s^2 found from their differences be worth more than 2 degrees of freedom? (In fact, it is often worth more nearly 1 df.) For intermediate order statistics (e.g. medians or quartiles) we can, indeed, improve the performance of our error estimation considerably by going to a different form of jackknife, to be discussed in a later section.

22. The simplest blocked - - or stratified - - jackknife.

If we have h blocks, labelled by i from 1 to h, and, in each block, m, we can apply the jackknife calculation to each block in turn, omitting, from that block, each piece in turn. The formulas are

$$y_{*ij} = m y_{ALL} - (m-1) y_{i(j)}$$

$$y_{*i} = \frac{1}{m} \sum y_{*ij}$$

$$s_{*i}^2 = \sum (y_{*ij} - y_{*i})^2 / (m-1)$$

$$y_{*.} = \frac{1}{h} \sum y_{*i}$$

$$hs_{\bullet}^2 = s_{\bullet 1}^2 + s_{\bullet 2}^2 + \ldots + s_{\bullet h}^2 = m(\textit{est'd var for } \textstyle\sum y_{\bullet i})$$

so that we may appropriately take

$$t = \frac{y_{\bullet} - \text{its contemplated value}}{\sqrt{s_{\bullet}^2/mh}}$$

with

$$s_{\bullet}^2 = \frac{1}{h}\sum_i s_{\bullet i}^2$$

If we have more than 2 pieces per block, we are likely to want to use this approach.

23. Jackknifing by halves.

The special case where $h=2$, when each block consists of 2 pieces, can be approached a little differently. Instead of "leaving out" half of a single block, going through the blocks in turn, we can "leave out" a half of each block in response to a preassigned sequence of two-fold choices. With two halves, "leaving out" one piece means "keeping" the other piece, and *vice versa*. Thus, if

"+" means 1st piece in *left*-hand half, 2nd in right-hand half

and

"−" means 1st piece in *right*-hand half, 2nd in left-hand half

Then the sequence + − + + divides 8 pieces (AB)(CD)(EF)(GH) - - in 4 blocks as follows - - into:

a left-hand half made up of ADEG, and

a right-hand half made up of BCFH

If g runs through some set of such sequences of +'s and -'s, and hence through some set of halvings, we can write:

y_{L_g} = y distilled from the left half of halving g

y_{R_g} = y distilled from the right half of halving g

The usual jackknife formulas for $m = 2$ now give

$$y_{*R_g} = 2y_{ALL} - y_{L_g}$$

$$y_{*L_g} = 2y_{ALL} - y_{R_g}$$

$$y_{*_g} = \frac{1}{2}y_{*R_g} + \frac{1}{2}y_{*L_g} = 2y_{ALL} - \frac{1}{2}(y_{L_g} + y_{R_g})$$

$$y_{*R_g} - y_{*_g} = \frac{1}{2}(y_{R_g} - y_{L_g})$$

$$y_{*L_g} - y_{*_g} = \frac{1}{2}(y_{L_g} - y_{R_g})$$

$$s^2_{*_g} = (y_{*R_g} - y_{*_g})^2 = (y_{*L_g} - y_{*_g})^2 = \frac{1}{2}(y_{R_g} - y_{L_g})^2$$

$$s^2_{*_g}/2 = \frac{1}{4}(y_{R_g} - y_{L_g})^2 = \textit{est'd var for } y_{*_g}$$

If now we average over g - - over our set of halvings - - we get, if there were G halvings:

$$y_* = 2y_{ALL} - \frac{1}{G}\sum_g \frac{1}{2}(y_{L_g} + y_{R_g})$$

$$= 2y_{ALL} - \text{mean of } y\text{'s for halves}$$

corresponding to a bias correction of

$$y_{ALL} - \text{mean of } y\text{'s for halvings}$$

which is appropriate for a bias proportional to 1/(amount of data) (since doubling the amount of data takes (1/amount of data) half-way to its value, 0, for an infinite amount of data.

Such averaging also gives

$$s_*^2 = \frac{1}{G}\sum_g \frac{1}{4}(y_{Lg}-y_{Rg})^2$$

Because many more different sets of pieces are left out, this "multi-halver" jackknife can acquire a reasonable number of degrees of freedom, even when jackknifing an intermediate order statistic.

24. How many, and which, halvings?

No matter how many halvings we manage to calculate, we cannot deserve more degrees of freedom than we started with, which was surely < mh. A counsel of perfection might call for halving according to all 2^h sequences of m +'s and −'s. But so much calculation would hardly be worthwhile, especially for substantial values of h. Yet the symmetry of doing all halvings is attractive.

We can obtain the desired symmetry with many fewer halvings by a simple device. Consider h=4 as an illustration. The 4 sequences (rows)

```
+  +  +  +
+  +  −  −
+  −  +  −
+  −  −  +
```

have a nice symmetric relation to one another, each pair (of rows) agreeing in two positions and disagreeing in two. The more interesting fact is that if we fix on any two columns, and look from halving to halving, these columns will show two matches and two disagreements - - a perfect balance.

One reason this fact is more interesting is that if there were only 3 blocks,

the set of truncated halvings

$$\begin{array}{ccc} + & + & + \\ + & + & - \\ + & - & + \\ + & - & - \end{array}$$

would still show this second sort of balance.

The four sign patterns we have been considering are relatively familiar, they define the usual factorial contrasts in 2^2 factorial experiment (where the grand mean is also of interest). There is no difficulty of extending all this to the 2^p contrasts (including the grand mean) for a 2^p factorial experiment. (And if it is important to use intermediate numbers of sequences we can go to Plackett-Burman designs.)

For the very nice case of an arithmetic mean, it is not hard to show that

$$s_*^2 = \frac{1}{G}\sum_g \frac{1}{4}(y_{L_g} - y_{R_g})^2$$

is exactly the appropriate multiple of the s^2 based on the individual y_i. For the "easy" cases then we can expect the multihalver to do essentially as well as possible. In the case of one year of data, with 3-day pieces and 2-piece (6-day) blocks, for instance we would have 61 blocks (trivially in leap years) and could use $p=6$ to generate $2^6 = 64$ sequences that were nicely balanced for 64 blocks, and that retained the important part of their balance for 61 blocks.

25. Stronger multihalver jackknives.

The property of the columns of the design matrix of +'s and -'s just discussed, that any *two* interrelate in a quite balanced way (once we realize that the complementary sequence refers to a transposed separation into the same

two halves). This says that the design matrix is an orthogonal array of strength two.

There are also arrays, somewhat larger (more rows, more sequences) of higher strengths, three, four, etc.

Orthogonal arrays of strength 3 for this problem are not difficult to use. Bose and Bush (1952) point out the existence of such patterns (sets of sequences) for halving up to 2^r blocks using only 2.2^r sequences (2.2^r halvings). This need never involve as many as 4 times as many halvings as blocks.

Orthogonal arrays of strength 4 or 5 tend to require many more halvings. Brillinger, Jones and Tukey (1978) give some selected examples, requiring:

for strength 4:

512 halvings for up to 23 blocks

1024 halvings for up to 32 blocks

2048 halvings for up to 63 blocks

for strength 5:

1024 halvings for up to 24 blocks

2648 halvings for up to 33 blocks

4096 halvings for up to 64 blocks

These may require as many as 60 (for strength 4 at 33 blocks) or 120 (for strength 5 at 35 blocks) times as many halvings as blocks.

These much stronger multihalvers are only likely to be needed in rather extreme circumstances.

26. Pros and cons of bias adjustment.

An important aspect of the jackknife is its bias-correction. This serves to allow, to a reasonable degree, both for known oversights in formulating the process by which a result is distilled from the data and for unrecognized dependencies on the amount of data.

We can choose to reduce the importance of bias-correction, both by how we formulate the process of distillation and how we choose to express our results - - see Section 28 below. We should strive to do the things as well as we can - - it is almost always better to avoid trouble than to try to fix it, but we should not expect to be perfect in avoiding this sort of bias - - some amount of bias decreasing like 1/(amount of data) is likely to remain. No matter how careful we are in formulation and expression, there is likely to be use for bias correction.

The nature of the jackknife's bias correction inhibits its use with very many pieces per block. The formulas for the pseudovalues

$$y_{*j} = my_{ALL} - (m-1)y_{i(j)} = y_{ALL} + (m-1)(y_{ALL} - y_{i(j)})$$
$$y_{*} = my_{ALL} - (m-1)y_{i(\cdot)} = y_{ALL} + (m-1)(y_{ALL} - \bar{y}_{i(\cdot)})$$

where $\bar{y}_{i(\cdot)} = (y_{i(1)} + y_{i(2)} + \ldots + y_{i(m)})/m$, with their factors of m and $m-1$ force the calculation of y_{ALL} and the $y_{i(j)}$ to additional precision. For m's between 5 and 20 this amounts to requiring about one extra decimal place or significant figure, not a great problem. (We are likely to be already keeping precision beyond that necessary for a not-bias-adjusted y_{ALL}, so that, for this range of m, it is often unnecessary to take special precautions.)

Much larger values of m can, and would, cause precision problems. Thus, much of our experience with the jackknife is in the range $4 \leq m \leq 20$.

We can always make m smaller by (a) using larger pieces and/or (b) using more blocks. Both of these reduce the number of degrees of freedom available - - (a) substantially and (b) slightly. But either or both, as we have already seen in the context of weather-related data, will often lead to the use of a more appropriate error term. It is more important to estimate a more appropriate σ^2 than it is to estimate whatever σ^2 we choose with somewhat more precision.

On balance, the existence of a bias correction in the jackknife helps us much more than it hurts us.

This view seems appropriate, even after we study the effect of bias correction on other types of bias, such as bias which decreases more rapidly, say like 1/(amount of data)^2. Such inverse quadratic biases come through the ordinary jackknifes - - leave-out-one or multihalver - - with an opposite sign and a somewhat reduced magnitude ($\times(1-\frac{1}{m})$ for leave-out-one, hence $\times(1/2)$ for multihalver).

There are higher-order jackknifes, (cp. book by Gray and Schucany, 1972), but they involve magnification factors of order m^2 and do not seem computationally acceptable.

27. Jackknife heuristics.

We have emphasized that the jackknife pseudovalues are intended to copy a sample *only* through second moments, to do a reasonably good job of making

$ave\{s_\bullet^2/m\} = var\{y_\bullet\}$. What about some more upbeat heuristics, at least so far as the structure of calculation for the pseudovalues and their analogs?

The basic need is to accommodate the existence of non-linearity - - doing this by finding modified values that can be well enough treated linearly.

The key process seems to involve:

1) a first step

2) turning the non-linearity loose

3) a final step

where, if there were no non-linearity at (2), the final step is so chosen as to undo the first step.

In the "leave-out-one" jackknife the first step involves leaving out one piece (or no pieces), the second step involves distilling a result according to the chosen algorithm, the third step involves the calculation of pseudovalues.

In the equally weighted case, the only one discussed explicitly in this report, *no non-linearity* would mean (some multiple of) forming an equally weighted mean, so that, in this special case,

$$y_{(i)} = \frac{-y_i + \sum y_j}{-1 + m} = \frac{m\bar{y} - y_i}{m-1}$$

$$y_{\bullet_i} = m\bar{y} - (m-1)y_{(i)} = (\sum y_j) - (\sum y_j) + y_i = y_i$$

(in greater generality, we could consider [some multiple of] a weighted mean

$$\bar{y}^{\,w} = \frac{\sum w_j y_j}{\sum w_j}$$

$$y_{(i)w} = \frac{-w_i y_i + \sum w_j y_j}{-w_i + \sum w_j} = \frac{(\sum w_j)\bar{y}^{\,w} - w_i y_i}{(\sum w_j) - w_i}$$

$$y_{*iw} = \frac{(\sum w_j)\bar{y}^{\,w} - ((\sum w_j) - w_i)y_{(i)w}}{w_i}$$

$$= \frac{(\sum w_j y_j) - (\sum w_j y_j) + w_i y_i}{w_i} = y_i$$

and thereby construct a weighted jackknife.)

The fact that the final step reverses the first step, when the intervening calculation is linear, is of course *not sufficient* to provide the useful properties of the jackknife. We need to follow through, at least for simple non-linearities, and see how well $ave\{s_*^2 1/m\}$ matches $var\{y_*\}$. But this simple pattern can at least help us remember the form of the jackknife calculation and, at best, give us a somewhat better "feel" for what is going on.

Something similar goes on with the "multihalver" jackknife, where the first step involves halving - - in which each block of two pieces is divided between left and right halves in accordance with the corresponding sign in a sequence of +'s and -'s. If we write E_{ig} for the i^{th} sign and the g^{th} sequence, we can say that we put piece P_{i1} into the left half $(1+E_{ig})/2$ times and into the right half $(1-E_{ig})/2$, putting its mate, P_{i2}, into the left half $(1-E_{ig})/2$ times and into the right half $(1+E_{ig})/2$ times.

In the very special case of linearity with equal weights, we could have

$$ky_{Lg} = \frac{1}{2}\sum(1+E_{ig})\bar{y}_{i1} + \frac{1}{2}\sum(1-E_{ig})\bar{y}_{i2}$$

$$= \frac{1}{2}\sum\bar{y}_{i1} + \frac{1}{2}\sum y_{i2} + \sum E_{ig}(\bar{y}_{i1} - \bar{y}_{i2})$$

$$= k\bar{\bar{y}} + \frac{1}{2}\sum E_{ig}(\bar{y}_{i1} - \bar{y}_{i2})$$

so that the corresponding pseudo-value

$$y_{*Lg} = 2\bar{\bar{y}} - y_{Lg}$$

$$= \bar{\bar{y}} - \frac{1}{2}\sum E_{ig}(\bar{y}_{i1}-\bar{y}_{i2})$$

whence

$$y_{\cdot L g} - y_{\cdot R g} = -\sum E_{ig}(\bar{y}_{i1}-\bar{y}_{i2})$$

and we can recover the $\bar{y}_{i1}-\bar{y}_{i2}$, by applying an inverse of $\{E_{ig}\}$. (If we use an orthogonal array of strength at least 2, $\{E_{ig}\}$ provides the necessary columns for an inverse.) We want to do this as a basis for (a) calculating $s_{\cdot}^2$ (which we could have done from the $y_{\cdot L g} - y_{\cdot R g}$) and (b) doing this with allowance for "seasonality of variability" - - for different $ave(\bar{y}_{i1}-\bar{y}_{i2})^2$ for different blocks (which requires getting hold of difference associated with blocks, not differences associated with sequences) - - so as to obtain a more appropriate number of degrees of freedom (cp. Section 10).

Here, so far as variance assessment goes, the first step is the halving according to each of several sequences, the middle step involves both distilling of each half and the comparison (by subtraction) of each pair of matched halves, and the final step undoes the linear transformation inherent in the halving.

Presumably other types of jackknife, involving the same three kinds of steps, are likely to appear in the future.

28. Care in the use of the jackknife.

If we jackknife y^2 instead of y - - where y is a well-specified distillate of the data - - we can expect somewhat inequivalent confidence intervals. A very simple case would involve $m=2$, $y_{(1)} = 4$, $y_{(2)} = 6$, $y_{ALL} = 5$, so that the pseudovalues will be $10-4 = 6$ and $10-6 = 4$. Hence, $y_{\cdot} = 5$, $s_{\cdot}^2 = (1^2 + 1^2)/1 = 2$

and

$$y_\bullet \pm 2\sqrt{(s_\bullet^2/2)} = 5 \pm 2 = \text{from 3 to 7}$$

If we jackknife $z = y^2$ instead, we would have $z_{(1)} = 16$, $z_{(2)} = 36$, $z_{ALL} = 25$, so that the pseudovalues will be $50-16 = 34$ and $50-36 = 14$. Hence, $z_\bullet = 24$, $s_{z\bullet}^2 = (10^2+10^2)/1 = 200$ and

$$z_\bullet \pm 2\sqrt{(s_{z\bullet}^2/2)} = 24 \pm 2\sqrt{100} = \text{from 4 to 44}$$

corresponding to an interval on the y scale of

$$\text{from } 2 = \sqrt{4} \text{ to } 6.6 = \sqrt{44}$$

clearly both ends of the confidence interval have been altered.

As noted above, one reasonable ground for choosing between jackknifing y and jackknifing y^2 - - if we have the necessary knowledge or insight - - would be to choose the expression whose sampling distribution is more nearly symmetrical.

Sometimes we can build in approximate symmetry through our choice of distillate. If we have two schemes, A and B, for predicting certain values that will be observed, if we are prepared to choose a measure of imperfection of prediction, of whatever reasonable form, and if we are concerned with comparing the two schemes, then we might reasonably plan to distill and jackknife

(imperfection of prediction of A)
MINUS
(imperfection of prediction of B)

which should tend to have a symmetrical distribution, at least in the null case where the two schemes produce predictions of equivalent quality (though

involving different values).

* second example *

Another sort of example may also be helpful. Suppose that we are distilling the second highest hourly value, for the whole year, of some weather-related value. (It is clearly easier to describe how we calculate such a distilled value than it is to define a parameter which it is supposed to estimate.) The failure of the bootstrap to have a bias correction (all "sample sizes" the same) implies no need to make a definition for any other number of observations than 24×365 or 24×366. For a leave-out-one jackknife we will need a definition for *a few days* (a few $\times$ 24 hours) *less than one year's* data. For a multihalver jackknife we will need a definition for approximately half a year of hours.

In this situation, we face, when we make our choices, not possible vs impossible, but better vs worse. We could define our general distilling process as "take the second highest of whatever values you have". This would mean relying on the bias correction to pick up the pieces. And there would be things for it to do! Since our half years, because of the blocks of the multihalver, are spread all through the year, the second largest from one half will be about the same size as the second largest from the complementary half, and thus will be, perhaps, the third, or fourth highest for the year as a whole.

This choice means (a) that we have thrown the difference between the second and the third-or-fourth highest (for the full year) into the bias correction, and (b) that we have forced ourselves to indicate that the target of

our distillation - - the corresponding parameter - - is the second highest value, even in a record of many, many years duration. It is unlikely that we really want either (a) or (b). While this choice of distilling process is possible, it is also both worse and unsatisfactory.

If instead we say we shall try to distill out a pointer to the $(2/24\times 365)n^{th}$ largest value (out of n candidates) then we will ask for almost the second highest for a "leave-out-one" jackknife, but for about the highest itself for each of the half-years used in our multihalver jackknife. Now we leave much less for our bias correction to do, and we indicate that, if we had 100 years of data, the 200^{th} highest value would satisfy us. These are both reasonable things, better than the previous choice.

This calculation made an assumption, one that we can improve slightly. It is that we should associate the i^{th} largest out of n with the fraction i/n - - an unsymmetric choice with $i=1$ giving $1/n > 0$, but $i=n$ giving $1 = 1$ not < 1. A variety of arguments suggest to me, not too strongly, but definitely, that it would be good (perhaps better) to assign the fraction $(3i-1)/((3n+1)$ to the i^{th} highest value. If we do this, the fraction $5/(3 \times 24 \times 365)+1)$ is assigned to the 2^{nd} highest for the year. The corresponding order statistic for the half year is j, where

$$\frac{(3j-1)}{((3\times 24\times 365/2)+1)} = \frac{5}{(3\times 24\times 365)+1}$$

or, nearly enough, the j where $(3j-1) = 5/2$ or $j = 7/6$, corresponding to interpolation about 1/6 of the way from the highest of the half-year toward the 2^{nd} highest for the same half year. (It would also correspond to taking about

the 167th highest out of a 100-year record.) I would expect this to be a slightly better choice than the previous one; others might not. Either will surely be better than the "always take the second highest" with which we began this example.

* regaining some degrees of freedom *

So much for bias and "what would we do with much more data". We usually have viable solutions. But what about minuscule numbers of degrees of freedom? Whatever bootstrap or jackknife we use, a "highest" or "second highest", leaves us poorly treated in this regard. Our need, then, is to find something which "points toward" a second highest, but whose value depends on many more observations. A reasonable class of possibilities are brought to mind by the words "exponential tail fit". If we are willing to assume a nominal shape for the extreme tail of the distribution of our values, thus redefining (we hope not by too much) what our distillate is pointing toward, we can fit a tail of the chosen shape - - perhaps to the highest 25 values - - and read off our distillate from the fitted curve at the point corresponding to the second highest value.

Doing something like this can earn us a fair number of additional degrees of freedom. The value we distill may - - or may not - - be somewhat more precise, but our estimate of its precision will be better, the critical values of Student's t will be smaller, the resulting confidence interval will be shorter.

Making such a change in our distillation process will give us a slightly modified target which we can estimate more closely. This will usually be a

good thing.

There is no reason not to combine this choice with the better dependence on the size of the body of data involved discussed just previously. Together they may help us quite a lot.

29. Summary of PART C.

Jackknives basically look at a finite "sample" of pieces as a finite population, and resample "subsamples" of different sizes. Blocking offers no particular difficulty. At one extreme the subsamples of a block include all but one of that block's pieces (the leave-out-one or basic jackknife). At another, blocks are of only two pieces each and the subsamples contain exactly half the data, one piece from each block. In this multihalver jackknife a number of halvings are used, perhaps a few more than there are blocks, sometimes several times this number, occasionally dozens of times this number.

The exact results of the jackknives depend on how we have expressed the distillate and its parameter, which ordinarily should be expressed in the same way. Choosing an expression is part of the user's responsibility, and allows us to do better than otherwise, ordinarily by seeking symmetry of distribution - - or translation-like behavior of the parameter - - or something between these two.

The bias correction built into the jackknives is usually helpful, particularly so when the expression is well-chosen.

Like the bootstrap, the jackknife is intended to help us with confidence intervals when distillates are not found by a simple adding up. Heuristically it

may help to think of a 3-step process, in which the 3rd step would undo the 1st step if the second step were not there.

PART D: Discussion

30. Many pieces per block.

If we have many pieces to our blocks, conventional bootstrap methods are a competitor for the jackknife. We recognize two main difficulties (and wish we knew if there are others):

1) the bootstrap invitably causes us to repeat 1/4 to 1/3 of all pieces in a typical resampling,

2) the naive bootstrap interval reacts to skewness in the wrong direction, in the desirable situation where the parameter is a location parameter. We have also to recognize the very limited character of available assessments of the rough size of the leading terms in the failure of the bootstrap to be exact. Asymptotics without some idea of error size could be, for all we know, a weak reed.

Either leave-out-one or multihalver (this one would require redefinition of blocks) jackknives avoid difficulty (1) and weaken (2) to:

(2*) Student's t produces symmetrical intervals, in whatever scale of expression is chosen - - thus we do have to choose an expression, but can gain by doing this well.

An argument could be - - and perhaps should be - - made for modifying

the naive bootstrap to deal more effectively with (2). Some possibilities (see also Section 19) include:

A) Using the bootstrap replications only to assess an estimated variance, and going over to Student's t (we would now, like the jackknife, have symmetric confidence intervals and the choice of expression would matter)

B) using reflected % points, of the replications, for instance

2(broadened median of bootstraps) − (upper 25% point of bootstraps)

as the lower end of the confidence interval,

2(broadened median of bootstrap) − (lower 25% point of bootstraps)

as the upper end. (The confidence interval would now depend more obviously on the scale in which the distillate is expressed. Probably dangerous for both small and moderate numbers of pieces per block.)

C) We might try to re-express the distillate - - perhaps through $y \to (e^{cy}-1)/c$ for a chosen c - - to promote symmetry of the bootstrap distribution, at least between the 1% and 10% points, and then go to (A).

31. Moderate numbers of pieces per block.

Until we have a better understanding which choices among the many modifications of the bootstrap that have the same asymptotic properties seem to lead to better performance, it is hard for me to urge the bootstrap as a serious competitor in this range of pieces per block. We do not know as much as we should about the jackknife in this range, but we know less about the bootstrap.

The "n" visible in all elementary discussions (usually for unstructured data) of the bootstrap will, in a blocked bootstrap, presumably fall somewhere between the number of pieces per block and that number multiplied by the number of *effective* (or *relevant)* blocks. (Ineffective, irrelevant blocks, like snow in summer months, cannot help us increase "n".)

32. Few pieces per block.

Here the multihalver jackknife - - and its possible generalizations to 3 or even 4 pieces per block - - seems to be the only reasonable candidate. It would be nice to know more about the properties of this approach, but, until we do, it is rather clearly best to steam right ahead and use the technique.

There could be a multiselector bootstrap, where the elements of our sequences describe the selection - - for example of 2 from 2 - - in each block. It seems unlikely that even an asymptotic justification could be found.

33. Differences in approach: bootstrap vs jackknife.

It may or may not have had to happen that the most traditional attitudes involved in approaching the bootstrap differ in flavor from the most traditional attitudes in approaching the jackknife. But they do differ - - and their difference illuminates the choice between traditional bootstraps and traditional jackknives.

The asymptotically sensible attitude common to most bootstrap approaches - - take care of everything that's of order $n^{-1/2}$ before taking anything of order n^{-1}, and so on - - is an advantage for sufficiently large n. (Recall that, for the

present account, n, which we will write "n" in this section, is something between the number of pieces per block and that number multiplied by the *effective* number of blocks.) How large is sufficiently large? It is quite plausible that 1000 is quite sure to be sufficient; it is implausible that 100 might be (at least most of the time); it is very doubtful that 10 will ever be sufficiently large.

The typical attitude surrounding the jackknife is quite different, having two major foundations:

1) If we know how to handle something, we do so (whether or not unhandleable things might be larger).

2) "Knowing" often has to be taken as depending on "n".

Studentizing - - referring to some one of Student's t distributions rather than to the unit Gaussian distribution - - is an illuminating instance. The only thing that has to be chosen is a number of degrees of freedom - - if this depends only on externals, not on the data, then we can Studentize quite well for small "n" (where degrees of freedom can be quite small).

If we have to estimate a number of degrees of freedom from the data, there will be some crossover "n_{xo}" such that we lose on average by Studentizing when "n" < "n_{xo}", but we gain if "n" > "n_{xo}". The crucial role in our thinking ought to be played by *crossovers* rather than by *asymptotic order*.

Instead of doing first everything of order $n^{1/2}$, and then things of higher asymptotic order, we plan to do those things that have "n" above the corresponding crossover. Thus, we will almost always Studentize, but will require a fairly large "n" before we start making the confidence interval

unsymmetrical around the (bias-adjusted) original distillate.

This latter approach has apparent advantages for smaller "n" - - it is plausible that these advantages are very real.

34. More on duplication in bootstrap replicates.

Bootstrapping special distillates that, so long as ties in the basic data are infrequent, are reasonable responses to plausible questions, can give rise to serious difficulty, because of the bootstrap's repetition of so large a fraction of pieces. If we do not expect many ties, and want to regard the data as a combination of a substantial number of small samples from different distributions of similar shapes and spreads, we may want to assess the (nearly common) spread of the constituent distributions. Arranging the observations in order of value and forming the gaps, the differences between adjacent values, followed by a summary of the lengths of the *shorter* gaps can be a reasonable approach to this question under these circumstances.

We might, for instance, distill:

a) the mean of the lower half of the gaps, or

b) the lower 33% value of the gaps.

Both of these will be distorted, to different degrees, by the additional ties generated by bootstrapping.

If there are several or many pieces per block, and we count repetitions in a weighted way (r appearances gets weight $r-1$) we find that there have to be additional ties covering an average of about 35%-37% of the gaps, thus

generating an equal % of additional zero gaps. This is enough to bias (a) very seriously - - and enough to force (b) to vanish for most bootstrap replicates.

While these examples are a little extreme, the extent of the failures suggest a need for serious worry about this problem, enough to be a further serious consideration favoring jackknives over bootstraps.

35. Improving the jackknife?

We have noted the possibility of using a weighted jackknife. The question of how best to generate the weight deserves some attention.

Considerations of third cumulants might be helpful, at least in cases where there are many degrees of freedom - - in either leave-out-one or multihalver forms.

Schemes for choosing an expression of the distillate that appear to have a reasonably symmetric distribution probably deserve attention. (It may be that this is a place of real usefulness for some form of bootstrap. Or it may be that we can make better use of the collective distribution of the distillates from all the halves of all the halvings considered.)

36. Basic philosophy.

I hope the flavor of this report will clearly distance me from those hypothetical people who "draw straight lines from inappropriate assumptions to inapplicable conclusions". If statistical techniques are to be helpful in practice, it will have to be because they work fairly well in real situations. A carpenter has to build a house using boards and timbers that have the shapes they have.

No one will be exactly a rectangular parallepiped. Building a data analysis needs to be thought of similarly - - as using what we have, for which no one assumption will hold exactly. It is our responsibility to understand, as best we can, how techniques work in real worlds. Knowledge of how they work in severely idealized worlds can help with this, but only if we take it for guidance, often quite rough guidance, and not for directly applicable truth.

37. Summary of PART D.

For many pieces per block, we can use the bootstrap, especially if we include a suitable subset of the modifications pointed to in Section 19.

For moderately many pieces per block we can happily use the leave-out-one jackknife; rather than have extremely many pieces per block we may want to reduce the block size somewhat.

For a moderate number of pieces per block, the writer would use a leave-out-one jackknife - - or change the block size to have only a few pieces per block.

For a few pieces per block, the writer would fatten pieces or shrink blocks to 2 pieces per block and then use a multihalver jackknife (ordinarily with only an orthogonal array (of halvings) of strength 2, exceptionally with one of strength 3 to 5.

There appear to be possibilities of improving both bootstraps and jackknives. Until these are tried out and understood, however, we should use the best that is currently available - - presumably that which is recommended at the start of this section.

This report has striven for a real world point of view.

REFERENCES

Bose, R. C. and Bush, K. A. 1952. "Orthogonal arrays of strength two and three," *Ann. Math. Statist.* 23: 508-524 (especially top of page 522).

Brillinger, D. R., Jones, L. V., and Tukey, J. W. 1978. "The role of statistics in weather resources management," *The Management of Weather Resources II.* (Report of the Weather Modification Advisory Panel) (especially page G-4).

Efron, B. 1987. "Better bootstrap confidence intervals," *J. Amer. Statist. Assoc.* 82: 171-185 (discussion 185-200, cp. also paper by DiCiccio and Tibshirani immediately preceding at 163-170).

Gray, H. L. and Schucany, W. R. 1972. *The Generalized Jackknife Statistic,* New York, M. Dekker.

END

10-87

DTIC